

Statistika

Testování hypotéz – statistická indukce – Neparametrické testy

Roman Biskup

(zapálený) statistik ve výslužbě, aktuálně analytik v praxi ;-)
roman.biskup(at)email.cz

21. února 2012



Obsah

Testování na základě jednoho výběru

Testování hypotézy o mediánu

Testování na základě dvou výběrů

Testování hypotézy o shodě dvou rozdělení

Testování na základě více než dvou výběrů

Testování hypotézy o shodě rozdělení

Testy mnohonásobného srovnávání



Znaménkový test I

Testování hypotézy o mediánu „spojitě“ rozděleného souboru

Nechť statistický znak X má v základním souboru přibližně rozdělení se spojitou distribuční funkcí

1. Hypotézy

$H_0:$		$H_A:$	K
$\tilde{x} \leq \tilde{x}_0$	$\tilde{x} > \tilde{x}_0$		$K = \{Y: Y \geq k_{2;\alpha}\}$ $K = \{u: u \geq u_{1-\alpha}\}$
$\tilde{x} = \tilde{x}_0$	$\tilde{x} = \tilde{x}_0$	$\tilde{x} \neq \tilde{x}_0$	$K = \{Y: Y \geq k_{2;\alpha/2} \vee Y \leq k_{1;\alpha/2}\}$ $K = \{u: u \geq u_{1-\alpha/2} \vee u \leq u_{\alpha/2}\}$
$\tilde{x} \geq \tilde{x}_0$	$\tilde{x} < \tilde{x}_0$		$K = \{Y: Y \leq k_{1;\alpha}\}$ $K = \{u: u \leq -u_{1-\alpha}\}$

2. Hladina významnosti: α

3. Volba testového kritéria: $Y = \#\{X_i: (X_i - \tilde{x}_0) > 0\}$

- ▶ kde $\#$ značí počet, konkrétně počet pozorování větších než $\tilde{x}_0 > 0$.
- ▶ $Y \sim \text{Bi}(n; 1/2)$ pro $n < 20$ za platnosti nulové hypotézy,
- ▶ $U = \frac{2Y-n}{\sqrt{n}} \sim U$ pro $n \geq 20$ za platnosti nulové hypotézy.

4. Vymezení kritického oboru: $K = \dots$



Znaménkový test II

Testování hypotézy o mediánu „spojitě“ rozděleného souboru

5. Výpočet testového kritéria: $y = \#\{x_i: (x_i - \tilde{x}_0) > 0\}$
6. Zjištění zda y nebo $u \in K$ a rozhodnutí: H_0 vs. H_A
7. Zformulovat slovní odpověď!

Poznámky:

- Neparametrická obdoba (jednovýběrového) t -testu.



Mannův-Whitneyův test (dvouvýběrový Wilcoxonův test) I

Testování hypotézy o shodě rozdělení dvou nezávislých souborů

Nechť statistické znaky X_1 a X_2 lze popsat například prostřednictvím jejich distribučních funkcí F_1 a F_2 .

1. Hypotézy: $H_0: F_1(x) = F_2(x)$ $H_A: F_1(x) \neq F_2(x)$

2. Hladina významnosti: α

3. Volba testového kritéria:
$$Z = \frac{T_1 - n_1(n_1 + n_2)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)}}$$

- ▶ kde T_1 je součet pořadí (celkového) pro statistický znak X_1
- ▶ $Z \sim N(0; 1)$ za platnosti nulové hypotézy + $\min(n_1; n_2) \geq 5$ (10; 30) (jinak tabelováno)

4. Vymezení kritického oboru: $K = \{Z: |Z| \geq u_{1-\alpha/2}\}$

5. Výpočet testového kritéria:
$$z = \frac{T_1 - n_1(n_1 + n_2)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)}}$$

6. Zjištění zda $z \in K$ a rozhodnutí: H_0 vs. H_A

7. Zformulovat slovní odpověď!

Poznámky:



Mannův-Whitneyův test (dvouvýběrový Wilcoxonův test) II

Testování hypotézy o shodě rozdělení dvou nezávislých souborů

- ▶ Obdobně možno spočítat Z přes T_2 (Pozor: $T_2 - n_2(n_1 + n_2)/2$). Z vyjde až na znaménko stejně. Pro oboustrannou hypotézu nehraje roli.
- ▶ Při výpočtu jsou obvykle vedle T_i vypočítávány i statistiky $U_i = T_i - \frac{n_i(n_i+1)}{2}$, testové kritérium pak lze získat jako $Z = \frac{U_i - n_1 n_2 / 2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)}}$.
- ▶ Tabelované hodnoty jsou určovány tak, aby hodnoty testových kritérií svědčící proti nulové hypotéze, tj. např. U_i blízko 0, či $n_1 n_2$, nastávaly s pravděpodobnostmi menší nebo rovnou α .
- ▶ p -value je tedy vypočteno jako pravděpodobnosti všech situací popírající nulovou hypotézu více nebo stejně než ukazují data, tj.
 $P(U \geq \max_i (U_i)) + P(U \leq \min_i (U_i))$, kde $U = 0, \dots, n_1 n_2$.
- ▶ Pokud není Mannův-Whitneyův test použit za předpokladu o shodě tvaru rozdělení pro test hypotézy $\tilde{x}_1 = \tilde{x}_2$ (pouze posunutí) je lépe použít Kolmogorovův-Smirnovův.
- ▶ Takto testové kritérium počítal Wilcoxon. Mann a Whitney počítali $\#\{[X_{1i}, X_{2j}]: X_{1i} < X_{2j}, i = 1, \dots, n_1, j = 1, \dots, n_2\}$.



Mannův-Whitneyův test (dvouvýběrový Wilcoxonův test)

III

Testování hypotézy o shodě rozdělení dvou nezávislých souborů

- ▶ Neparametrická obdoba dvouvýběrového t -testu – zvláště pokud se jedná o posouzení posunutí (v tomto případě nejsilnější).
- ▶ Pojmenováno podle HENRYho BERTHOLDa MANNa (1905–2000) a DONALDa RANSOMA WHITNEYe (1915–2001) respektive FRANKa WILCOXONa (1892–1965).



Mannův-Whitneyův test – test o existenci posunutí

Testování hypotézy o shodě parametru úrovně rozdělení dvou nezávislých souborů

Nechť statistické znaky X_1 a X_2 lze popsat například prostřednictvím jejich distribučních funkcí F_1 a F_2 , které mají přibližně stejný tvar.

1. Hypotézy

H_0 :	H_A :	K
$\tilde{x}_1 \leq \tilde{x}_2$	$\tilde{x}_1 > \tilde{x}_2$	$K = \{Z: Z \geq u_{1-\alpha}\}$
$\tilde{x}_1 = \tilde{x}_2$	$\tilde{x}_1 \neq \tilde{x}_2$	$K = \{Z: Z \geq u_{1-\alpha/2}\}$
$\tilde{x}_1 \geq \tilde{x}_2$	$\tilde{x}_1 < \tilde{x}_2$	$K = \{Z: Z \leq -u_{1-\alpha}\}$

2. Hladina významnosti: α

3. Volba testového kritéria: $Z = \frac{T_1 - n_1(n_1 + n_2)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)}}$

- ▶ kde značení je stejné jako u běžného Mannova-Whitneyova testu
- ▶ $Z \sim N(0; 1)$ za platnosti nulové hypotézy + $\min(n_1; n_2) \geq 5$ (10; 30) (jinak tabelováno)

4. Vymezení kritického oboru: $K = \dots$

5. Výpočet testového kritéria: $z = \frac{T_1 - n_1(n_1 + n_2)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)}}$

6. Zjištění zda $z \in K$ a rozhodnutí: H_0 vs. H_A

7. Zformulovat slovní odpověď!



Dvouvýběrový Kolmogorovův-Smirnovův test I

Testování hypotézy o shodě rozdělení dvou nezávislých souborů

Nechť statistické znaky X_1 a X_2 lze popsat například prostřednictvím jejich distribučních funkcí F_1 a F_2 .

1. Hypotézy: $H_0: F_1(x) = F_2(x)$ $H_A: F_1(x) \neq F_2(x)$

2. Hladina významnosti: α

3. Volba testového kritéria: $D = \sup_x |F_1^e(x) - F_2^e(x)|$

- ▶ $F_1^e(x) = \frac{1}{n_1} \# \{X_{1i}: X_{1i} \leq x, i = 1, \dots, n_1\}$, empirická distribuční funkce pro statistický znak X_1 ,
- ▶ $F_2^e(x) = \frac{1}{n_2} \# \{X_{2i}: X_{2i} \leq x, i = 1, \dots, n_2\}$, empirická distribuční funkce pro statistický znak X_2 ,
- ▶ $D \sim ??$ za platnosti nulové hypotézy + $(n_1 + n_2) > 35$ (jinak tabelováno)

4. Vymezení kritického oboru: $K = \{\}$

5. Výpočet testového kritéria: $d = \sup_x |F_1^e(x) - F_2^e(x)|$

6. Zjištění zda $d \in K$ a rozhodnutí: H_0 vs. H_A

7. Zformulovat slovní odpověď!



Dvouvýběrový Kolmogorovův-Smirnovův test II

Testování hypotézy o shodě rozdělení dvou nezávislých souborů

Poznámky:

- ▶ Co může říkat alternativní hypotéza?:
 1. Rozdělení X a Y se liší svojí polohou (střední hodnotou, kvantily), přičemž jak variabilita tak tvar obou rozdělení mohou být shodné.
 2. Rozdělení X a Y se liší svojí variabilitou, přičemž jak poloha tak tvar obou rozdělení mohou být shodné.
 3. Rozdělení X a Y mají odlišný tvar, přičemž jak poloha tak variabilita obou rozdělení mohou být shodná.
- ▶ Pojmenováno podle ANDREJE NIKOLAJEVIČE KOLMOGOROVA (1903–1987) a IGORA NIKOLAEVICHE SMIRNOVA (1941–).



Kruskalův-Wallisův test I

Testování hypotézy o shodě rozdělení více jak dvou nezávislých souborů

Nechť statistické znaky $Y_1, Y_2, \dots, Y_k$ lze popsat například prostřednictvím jejich distribučních funkcí $F_1, F_2, \dots, F_k$.

1. Hypotézy: $H_0: F_1(y) = F_2(y) = \dots = F_k(y)$ $H_A: \text{non}H_0 (\neg H_0)$

2. Hladina významnosti: α

3. Volba testového kritéria: $H = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{T_i^2}{n_i} - 3(n+1)$

$$\blacktriangleright n = \sum_{i=1}^k n_i$$

$\blacktriangleright T_i$ je součet pořadí (celkového) pro statistický znak Y_i

$\blacktriangleright H_{kor} = \frac{H}{1 - \frac{\sum_{p=1}^q (t_p^3 - t_p)}{n^3 - n}}$ – korekce stejného pořadí, kde t_p je počet stejných

průměrných pořadí, $p = 1, 2, \dots, q$; q je počet variant průměrných pořadí.

$\blacktriangleright H, H_{kor} \sim \chi^2(k-1)$ za platnosti nulové hypotézy + $k \geq 4$, $n_i \geq 5$ (pro $k = 3$, $n_i < 5$ tabelováno)

4. Vymezení kritického oboru: $K = \{H: H \geq \chi_{1-\alpha}^2(k-1)\}$



Kruskalův-Wallisův test II

Testování hypotézy o shodě rozdělení více jak dvou nezávislých souborů

5. Výpočet testového kritéria: H a H_{kor}
6. Zjištění zda $H \in K$ a rozhodnutí: H_0 vs. H_A
7. Zformulovat slovní odpověď, v případě platnosti H_A pokračovat v analýze pro jednotlivé soubory!

Poznámky:

- Pojmenováno podle WILLIAMA HENRYho KRUSKALA (1919–2005) a ALLENA WILSONA WALLISE (1912–1998).



Friedmanův test I

Testování hypotézy o shodě rozdělení více jak dvou závislých souborů

Nechť statistické znaky $Y_1, Y_2, \dots, Y_k$ lze popsat například prostřednictvím jejich distribučních funkcí $F_1, F_2, \dots, F_k$.

1. Hypotézy: $H_0: F_1(y) = F_2(y) = \dots = F_k(y)$ $H_A: \text{non}H_0 (\neg H_0)$

2. Hladina významnosti: α

3. Volba testového kritéria: $Q = \frac{12}{nk(k+1)} \sum_{i=1}^k T_i^2 - 3n(k+1)$

► kde T_i je součet pořadí (v rámci k -tice) pro statistický znak Y_i

► $Q_{kor} = Q \left[\frac{n}{n - \frac{1}{k^3 - k} \left(\sum_{p=1}^q \sum_{s=1}^{r_p} (t_{ps}^3 - t_{ps}) \right)} \right]$ – korekce stejného pořadí, kde

t_{ps} je počet stejných průměrných pořadí u p -tého objektu (měl-li průměrné pořadí), $p = 1, 2, \dots, q$, $q < n$; s je počet stejného průměrného pořadí u téhož objektu, $s = 1, 2, \dots, r_p$; r_p je počet variant průměrných pořadí u p -tého objektu.

► $Q, Q_{kor} \sim \chi^2(k-1)$ za platnosti nulové hypotézy + $k \geq 3$, $n \geq 10$; nebo $k \geq 4$, $n \geq 5$; nebo $k \geq 5$, n libovolné (jinak tabelováno)



Friedmanův test II

Testování hypotézy o shodě rozdělení více jak dvou závislých souborů

4. Vymezení kritického oboru: $K = \{Q: Q \geq \chi_{1-\alpha}^2(k-1)\}$
5. Výpočet testového kritéria: Q a Q_{kor}
6. Zjištění zda $Q \in K$ a rozhodnutí: H_0 vs. H_A
7. Zformulovat slovní odpověď, v případě platnosti H_A pokračovat v analýze pro jednotlivé soubory!

Poznámky:

- Pojmenováno podle MILTONA FRIEDMANA (1912–2006).



Neményho test mnohonásobného srovnávání

Následné testování hypotézy o shodě rozdělení dvou závislých/nezávislých souborů

Nechť statistické znaky $Y_1, Y_2, \dots, Y_k$ lze popsat například prostřednictvím jejich distribučních funkcí $F_1, F_2, \dots, F_k$.

1. Hypotézy: $H_0: F_i(y) = F_j(y)$ $H_A: F_i(y) \neq F_j(y)$, pro $i, j = 1, \dots, k$, $i \neq j$
2. Hladina významnosti: α
3. Volba testového kritéria: $|T_i - T_j|$
 - ▶ T_i je součet pořadí vypočítaný v rámci Kruskal-Wallisova testu respektive Friedmanova testu
4. Vymezení kritického oboru: $K = \{|T_i - T_j| \geq N_{1-\alpha}(k, n)\}$
 - ▶ Jiné hodnoty pro nezávislé a závislé soubory, tj. jestli počítáme s T_i vypočtenými při Kruskal-Wallisově testu respektive Friedmanově testu!
5. Výpočet testového kritéria: $|T_i - T_j|$
6. Zjištění zda platí nerovnost a rozhodnutí: H_0 vs. H_A
7. Zformulovat slovní odpověď.

Poznámky:

- ▶ Pojmenováno podle PETERA NEMÉNYIHO (1927–2002).

